

FIG. 1

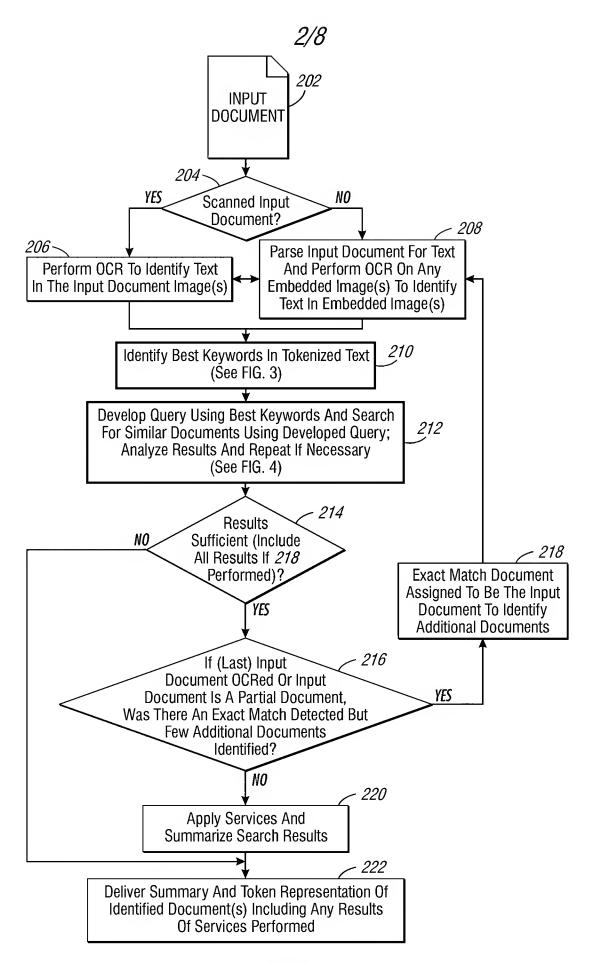


FIG. 2

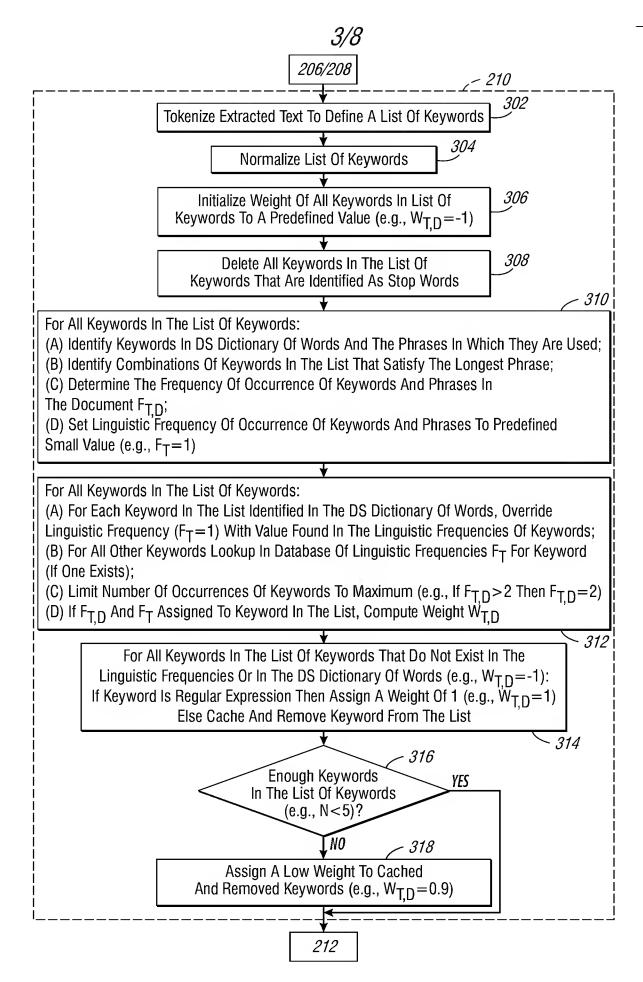


FIG. 3

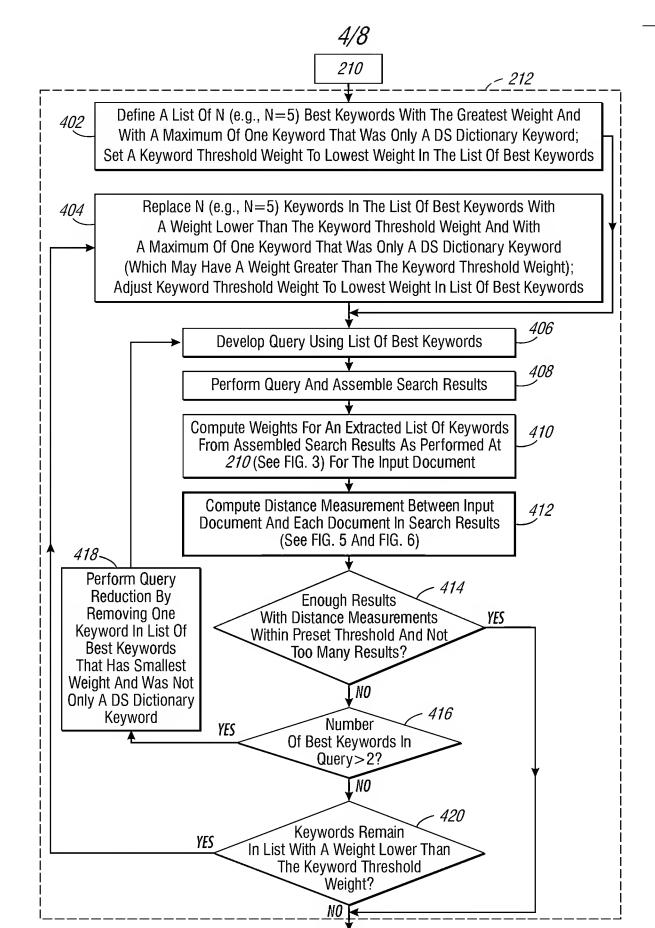


FIG. 4

214

CALCULATE_SIMILARITY [D1,D2] -calculates the similarity between documents D1 and D2

- mdur **,**
- D1: List of keywords of the input document
- D2: List of keywords of document from search results
- Output:
- Similarity "S": The computed similarity between D1 and D2

{ For the document D1, calculate with the keyword weights of D1:

- unique attributes sum: Sum1 = the sum of the weights of keywords in D1 that do not appear in D2
- total sum: Sum2 = the sum of the weights of keywords in D1
- shared sum: Sum3 = the sum of the weights of keywords in D1 that also appear in D2
- ratio: R = (the number of keywords in D1 not in D2)/(the number of keywords in D1)

| For the document D2, calculate with the keyword weights of D2:

- unique attributes sum: Sum4 = the sum of the weights of keywords in D2 that do not appear in D1
 - shared sum: Sum5 = the sum of weights of keywords in D2 that also appear in D1

If D1 originates from a hardcopy document, calculate the tolerance ratio "T":

K = a constant defined by OCR error rate at the keyword level, if no OCR error is detected, K is set to 0 $T = K^* (Sum2 - Sum3)/(Sum2)$

Calculate the inclusion ratio "I" (i.e., percentage of keywords from D1 that are in D2):

I = (Sum3)/(Sum2) + T

F/G. 5

```
FROM FIG. 5
                                If I>90% (i.e., if an inclusion is detected, e.g., 90 % of the keywords from D1 are in D2):
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             • Sum8 = Ordered_Sum [D2,D1] - sum of the weights of keywords from D2
                                                                                                      with same neighbors in D2 (SEE FIG. 7)
                                                                 Sum6 = Ordered_Sum [D1,D2] - sum of the weights of keywords in D1
                                                                                                                                                                                                                                                                                       if (D1 originates from a hardcopy document and I2>50%)
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    with same neighbors in D1
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    • If S>90%: (a revision is detected, i.e., Jaccard similarity S>90 %;
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           otherwise a related document may be detected)
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              if (D1 originates from a hardcopy document) if (S2>50%)
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     S2 = (Sum8) / (Sum1 + Sum4 + Sum8)
                                                                                                                                                                                                                                                                                                                                                                                                                                         · Calculate the Jaccard similarity distance measure:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       if (R>20%) then S = S1
                                                                                                                                                                                                                                                                                                                               if (R<20 %) then S = I else S = I2 else S = I2
                                                                                                                                       • Calculate ordered inclusion ratio "12":
                                                                                                                                                                                                                                                                                                                                                                                                                                                                             Sum7 = Sum1 + Sum4 + Sum5
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    • Calculate ordered similarity:
                                                                                                                                                                                                                                                                                                                                                                                                    else (i.e., if no inclusion is detected):
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             else S = S1
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                S = (Sum5) / (Sum7)
                                                                                                                                                                               I2 = (Sum6)/(Sum2)
                                                                                                                                                                                                                   if (I2>I) then S = I
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              else S = S1
                                                                                                                                                                                              510
```

FIG. 6

Ordered_Sum [L1, L2] - calculates sum of the weights of keywords from L1 with same neighbors in L2 increase the ordered sum with the weight of term t in L1: Sum += Wt if the position Pi of the term t is at a limit of L2: (Pi<N or Pi>(L2 size-N)) (by default N=5 and depends on the position of the term t in the L1) Define the tolerance "T" minimal percentage for neighbors: $\frac{6000}{1} = K*50\%, \text{ where K depends on the OCR error at the keyword level}$ identify N neighbors on both sides of term t in L1 identify all possible positions Pi of term t in L2 for each position Pi of term t found in L2: for each term t (i.e., keyword) of L1: • Sum: the ordered sum if (t exists in L2) • L1: List 1
• L2: List 2 Output

FIG. 7

if (C > 80% - T) increase the ordered sum with the weight of term t in L1: Sum += Wt

increase the ordered sum with the weight of term t in L1: Sum += Wt

Calculate the percentage of common neighbors "C" of term t between L1 and L2

identify N neighbors on both sides of term t in L2

